ANNUAL REVIEWS

*Annual Review of Genomics and Human Genetics*

# The UK Biobank: A Shining Example of Genome-Wide Association Study Science with the Power to Detect the Murky Complications of Real-World Epidemiology

Vanessa Y. Tan[1,2] and Nicholas J. Timpson[1,2]

[1]Medical Research Council (MRC) Integrative Epidemiology Unit, University of Bristol, Bristol, United Kingdom; email: n.j.timpson@bristol.ac.uk

[2]Bristol Medical School, University of Bristol, Bristol, United Kingdom

Annu. Rev. Genom. Hum. Genet. 2022. 23:569–89

First published as a Review in Advance on May 4, 2022

The *Annual Review of Genomics and Human Genetics* is online at genom.annualreviews.org

https://doi.org/10.1146/annurev-genom-121321-093606



## Keywords

UK Biobank, genome-wide association study, genetics

## Abstract

Genome-wide association studies (GWASs) have successfully identified thousands of genetic variants that are reliably associated with human traits. Although GWASs are restricted to certain variant frequencies, they have improved our understanding of the genetic architecture of complex traits and diseases. The UK Biobank (UKBB) has brought substantial analytical opportunity and performance to association studies. The dramatic expansion of many GWAS sample sizes afforded by the inclusion of UKBB data has improved the power of estimation of effect sizes but, critically, has done so in a context where phenotypic depth and precision enable outcome dissection and the application of epidemiological approaches. However, at the same time, the availability of such a large, well-curated, and deeply measured population-based collection has the capacity to increase our exposure to the many complications and inferential complexities associated with GWASs and other analyses. In this review, we discuss the impact that UKBB has had in the GWAS era, some of the opportunities that it brings, and exemplar challenges that illustrate the reality of using data from this world-leading resource.

## INTRODUCTION

Genetic architecture is defined by the genetic variants that influence a trait or disease outcome and is characterized by the number of genetic variants, their effect size, their allele frequency, and their possible interactions with each other and environmental factors (114). Uncovering the genetic architecture of a complex trait or disease is central to understanding what underpins observed trait variation and potentially helps further work aimed at dissecting that variation. Various techniques are useful in the task of assessing genetic architecture, and in the last decade, technological advances have enabled the measurement of genetic variation at hundreds of thousands of markers across the human genome. Genome-wide association studies (GWASs) have exploited these developments and successfully identified thousands of genetic variants reliably associated with human traits. Although GWASs are restricted to certain variant frequencies, they have improved our understanding of the genetic architecture of complex traits and diseases (118). Early successful applications of the GWAS approach have primed the rapid growth and increasing availability of large genomic biobanks, such as the UK Biobank (UKBB) (15, 109), which have brought substantial analytical power to association studies.

## THE REVOLUTIONARY IMPACT OF THE UK BIOBANK IN THE GWAS WORLD

UKBB (**https://www.ukbiobank.ac.uk**) is a prospective cohort study that recruited half a million individuals (40–69 years of age) across the United Kingdom between 2006 and 2010 (109). It represents a large-scale biomedical resource that integrates genome-wide genetic data with deep phenotype data, including data from lifestyle questionnaires, physical measures, biomarkers in blood and urine, accelerometry, and multimodal imaging. The unprecedented size of the UKBB cohort, together with the extensive phenotyping and genome-wide genotype data (supplemented with high-density imputation), has enhanced power for genetic discovery and enabled well-powered GWASs of hundreds of quantitative traits, including anthropometric traits (123), blood traits (4), cognitive traits (42), and numerous blood and urine biomarkers (105). Further to this enormous collection of data, the access infrastructure provided with the UKBB study has led to UKBB being one of the most enabling human genetics bioresources ever generated. To date, more than 500 peer-reviewed GWASs have been published based on this one resource, and UKBB is often the primary data provider in other analyses.

In the context of the capacity and coverage of GWAS arrays (and imputation), sample size is one of a series of key performance determinants when performing a GWAS to discover genetic loci associated with complex traits (118). This has been evident since the first application of hypothesis-free approaches to gene variant analysis (12, 77, 108), and (depending on the composition of heritable contributions) the discovery of new loci has increased in an almost linear fashion with increasing sample size (83). The availability of genome-wide genotype data collected from all UKBB participants, together with the biobank's vast amount of phenotype data, has generated a singular resource of considerable size that provides opportunities for the discovery of new genetic associations and the genetic basis of complex traits and diseases (15). The gain in power in the UKBB data is exemplified by the most recent meta-analysis of GWASs of height and body mass index (BMI) (123), which combined results from a single large GWAS conducted using UKBB data with results from previously published GWASs of height and BMI conducted by the Genetic Investigation of Anthropometric Traits (GIANT) study. With the increased sample size, the number of genomic loci associated with height and BMI increased compared with the number found in previously published height and BMI GWASs, with improved accuracy of genetic predictors from single-nucleotide polymorphisms (SNPs) at these loci (123). The near-independent

SNPs explained approximately 24.6% and 6% of the variance in height and BMI, respectively, representing approximately 1.9- and 3.2-fold improvements, respectively, in comparison with a previous BMI GWAS meta-analysis that did not include UKBB data (123).

The combination of scale and extensive phenotypic and genotypic data from UKBB enables the rigorous investigation of the genetic basis of diseases, not only through the increased sample size but also through phenotypic precision (66). Although expanding collections of genotype and phenotype data from non-UKBB studies have provided a boost in statistical power, research has been hindered by measurement differences, inaccurate phenotypic measurements, and genuine disease heterogeneity. One challenge in GWASs is the ability to combine genetic data with phenotypic precision and hence to both enhance analytical power technically and tighten the focus of downstream interpretation of findings around pertinent association signals (106, 116). In UKBB, the combination of deep phenotypic data collection and large-scale genetic data has generated biological insights and extensive records of novel genotype–phenotype associations. A good example is the use of detailed brain imaging data to examine the genetic architecture of brain structure in a GWAS of more than 3,000 functional and structural brain imaging phenotypes in more than 8,000 UKBB participants (27); this work showed not only that many of these phenotypes are likely to be heritable but also that they lie in phenotypic clusters showing reliable genetic associations. Another example is a study by Aragam et al. (3), who used UKBB data to perform a GWAS of heart failure and found that phenotypic refinement of all-cause heart failure facilitates the discovery of novel genetic signals that reflect distinct etiological heart failure subtypes.

## OVERLAPPING PHENOTYPES AND HARNESSING THE PHENOME

Most GWASs analyze only a single trait and do not exploit information on summary statistics from GWASs of other correlated traits. UKBB is clearly different in terms of the breadth of measures available across participants and the repeated assessments during the more than 10 years the study has been running. Uniquely, this has led to a situation where previously and newly developed approaches exploiting the information in shared phenotypic measures can be deployed and share the sample size advantages present in UKBB. Joint association analysis of multiple traits in GWASs offers several advantages compared with single-trait analyses and has been a well-used approach in the undertaking of GWASs in UKBB. Multivariate analysis can boost statistical power because it takes into account cross-trait covariance of genetically correlated traits, which is often ignored in univariate analyses (1, 135). Multivariate methods can also assess associations with a set of traits using a single test. Consequently, the multiple-testing burden can be reduced because of the effectively smaller number of analyses performed (55, 135). Lastly, when a single genetic variant is highly pleiotropic and associated with multiple traits, multivariate GWAS analysis can attempt to model a situation that is more consistent with the underlying biology (i.e., accounting for shared variance across multiple available measures with variable levels of actual information content across key heritable contributions) when compared with univariate analysis (18).

Several techniques have been developed that enable joint analysis of multiple traits (8, 34, 45, 76, 86, 115), and their application is one of the hallmarks of well-performed UKBB-based GWASs that seek to maximize analytical performance. A good example is Multi-Trait Analysis of GWAS (MTAG), which allows the joint analysis of multiple traits in population-based GWASs, thereby increasing the statistical power to detect genetic associations for each analyzed trait. Using MTAG, a recent study conducted a joint GWAS analysis of four hearing-related traits using UKBB data and identified 31 new risk loci for hearing difficulty (49).

## APPLIED ANALYSES ENABLED BY A GWAS BACKBONE

One of the most intuitive applied analyses built on the success of well-performed and well-powered GWASs is the examination of polygenic risk scores (PRSs) and the association of aggregate genetic variant effects with intermediates or outcomes of interest. These composite summaries of genetic association can vary in their form, ranging from the sum of association estimates across loci believed to have attained specific thresholds of evidence to genome-wide predictors that look to amass all information relevant to the prediction of specific health outcomes. Several studies have had limited success in obtaining meaningful predictive power (72, 96) in undertaking this type of exercise, particularly because of the difficulty of generating meaningful predictions of phenotypic value through the amalgamation of GWAS signals. Previous efforts to create effective PRSs have been limited by three challenges (53): (*a*) small GWAS sample sizes, which affect the precision of estimated variant effects and the extent and availability of the useful variance explained; (*b*) limited methods for creating the PRSs; and (*c*) a lack of large data sets to validate and test PRSs. In the context of the increasing number of GWASs (much of it driven by UKBB), studies have started to overcome some of these difficulties and to generate interesting and informative polygenic summaries. A recent study by Khera et al. (54) used UKBB data as a validation data set to test the ability of BMI PRSs to predict measured BMI. Using UKBB, this study demonstrated the ability to use PRSs to identify individuals at greatest risk of obesity, with more than 40% of individuals who had a PRS in the top decile found to be obese, compared with 10% of individuals in the bottom decile.

Mendelian randomization (MR) is an analytical technique that uses genetic variants as instruments to estimate the causal effect of an exposure on an outcome of interest (24). By exploiting the properties of genetic data, MR analyses provide an alternative source of evidence when estimating causal effects and attempting to minimize limitations through confounding, bias, and reverse causation. MR analyses can be undertaken by using individual studies with exposures, outcomes, and genetic data, but also by using the results from existing GWASs (58, 131). MR-Base (40)—an established and freely accessible online platform that combines a database of GWAS results with an interface for performing MR and sensitivity analyses—has simplified the implementation of MR studies and enabled users to explore millions of potentially causal associations. The expansion of large-scale GWASs using UKBB data has rapidly increased the number of genetic variants that have been reliably associated with human characteristics and health conditions. For example, since the release of the UKBB genome-wide association data, the Neale laboratory (79) has conducted GWASs for thousands of phenotypes. UKBB data have also been incorporated into the Medical Research Council (MRC) Integrative Epidemiology Unit (IEU) OpenGWAS database (28), an open source, open access, scalable, and high-performance cloud-based data infrastructure that imports and publishes complete GWAS summary data sets and metadata for the scientific community. Taken together, this open GWAS data resource and the development of results-based MR applications and newly available analytical tools have enabled causal inference analysis (28). For example, by using the IEU OpenGWAS database to obtain genetic instruments for body composition measures (including fat mass, fat-free mass, and fat percentage) from UKBB-based GWASs, a recent MR study found evidence that high fat mass and fat percentage causally increased risk of most cardiometabolic diseases, while high fat-free mass had protective effects on cardiometabolic diseases after accounting for fat mass (126).

A different but similarly applied approach that uses GWAS analysis and results is the examination of shared genetic architecture or genetic correlation. Genetic correlation is a quantitative parameter that estimates the shared heritable contribution to two traits. Identifying genetic correlations between complex traits and diseases can provide useful insights into disease etiology, help identify potential causal relationships (11), and increase understanding of shared biological

contributions to apparently independent traits. Methodological approaches that estimate SNP-based heritability and genetic correlations from GWASs, such as linkage disequilibrium score regression (LDSR) (11), have proven to be powerful tools that can provide robust estimates of the genetic correlations between different traits and diseases and help to dissect the genetic architecture of common traits and diseases. LDSR relies on GWAS summary statistics and is not biased by sample overlap, and thus is invaluable in increasing our knowledge of the genetic contribution to complex traits. A major challenge preventing accurate estimation of genetic correlation is that GWASs with small effective sample sizes have insufficient power to use LDSR to detect polygenic effects, leading to near-zero estimates of heritability. Recently, UKBB data have been used to accurately estimate the SNP heritability of 22 complex traits and disease traits (44) and the genetic correlations between various traits and diseases (80). An LDSR analysis of more than 2,000 phenotypes in UKBB found that substantial variance was explained by common SNPs for a broad range of human traits and diseases (78).

## ENHANCING THE UK BIOBANK RESOURCE AND THE DEMOCRATIZATION OF DATA

Given the prospective nature of the UKBB study, a key strength of the cohort is the collection of data, biosamples, and exposures at baseline, which can be enhanced by other data sets that may be available from external sources but are linked to the study. Electronic health records, which can be a key source of data that are not affected by attrition, are a substantial research asset and further illustrate the positive civic relationship between study and participants that lies at the heart of UKBB. These linkable resources include death and cancer registries and primary and secondary care records, and in the absence of attrition, there is potential to follow up on the health of all participants over time. Connection to health outcome data provides opportunities to conduct research on common diseases, such as ischemic heart disease and various cancers, and to further expand the portfolio of GWASs embedded in the UKBB resource. This powerful design also enables conditions that are difficult to study retrospectively, including dementia and rapidly fatal conditions such as pancreatic or lung cancer. For example, a recent pan-cancer GWAS provided insights into the complex genetic architecture of cross-cancer susceptibility by using linked cancer registry data from UKBB and the Kaiser Permanente Genetic Epidemiology Research on Adult Health and Aging cohort (92).

UKBB is an open access resource that encourages researchers from around the world—including those from both academia and industry—to access the data and biological samples to undertake health-related research that is in the public interest. The open access nature of the UKBB study promotes innovative science by enabling international scientists to apply for the data quickly and easily through an application process so that they can benefit from this vast resource (19). Recently, to accommodate the vast scale of the data set, UKBB has launched the unique and innovative Research Analysis Platform, a cloud-based system that allows streamlined access for approved researchers from anywhere in the world and enables data to be analyzed easily and cost-effectively as the resource grows in complexity and scale. The open access data have not only enabled thousands of specific research projects but also encouraged collaborations with large international consortia. In the field of human genetics, and GWASs in particular, this type of data use and collaborative interchange is critical to building the sample sizes and measure/outcome capture needed to uncover reliable association signals. Ultimately, this approach to open and collaborative science has resulted in rapid advances in the discovery of reliable relationships and the beginnings of unrestricted researcher contributions to the task of dissecting the genetic architecture of complex traits.

## THE REALITY OF GWASs: POWER, POLYGENICITY, SAMPLING FRAME, AND INTERPRETATION

UKBB is an outstanding example of the value that can be achieved from large sample sizes combined with genetics, extensive and deep phenotyping, and linkage to health records. The gain in power in the UKBB cohort is clear and has led to an increase in locus discovery in GWASs, particularly for loci that are less common and/or have smaller effects (118). For example, the first BMI GWAS ($n \sim 5{,}000$) identified only genetic variants in the *FTO* locus that had relatively large effects on BMI (0.35 kg/m$^2$ per allele) (31, 102). By contrast, the most recent BMI GWAS, which used data from UKBB and the GIANT consortium ($n \sim 800{,}000$), identified more than 750 loci with much smaller effects on BMI (0.04 kg/m$^2$ per allele) (123). Notably, UKBB data represented 64.3% of the overall sample size. In this type of work, it is clear that the unprecedented size of the UKBB resource has provided immense opportunities; however, it can also generate challenges. To focus on just two of these that are relevant to GWASs (particularly applied GWAS work), here we examine the potential for population stratification/substructure to be important in the presence of specific GWASs undertaken at scale and the potential for detectable phenotypic overlap to complicate downstream interpretation and analysis.

Along with the potential to underrepresent specific groups and reduce generalizability (52), the self-selection of participants contributing to the UKBB cohort creates structure within the genetic data that has the potential to bias associations and complicate their interpretation. Although UKBB was designed to provide a large and useful sample of the general population of the United Kingdom (which it does superbly), the sampling population is volunteer based and not representative in terms of demographic characteristics (33). Ultimately, UKBB is a selected sample of the UK population (response rate of 5.5%) (112). This has been illustrated in work by Haworth et al. (39), who showed that single genetic variants and genetic scores composed of multiple variants are associated with the birth locations of individuals included in the UKBB data and that the geographic structure in genotype data cannot be accounted for by using routine adjustment for study center and genetic principal components. This study also demonstrated that major health outcomes appear to be geographically structured and that the coincident structure in health outcome and genotype data can yield biased associations.

As described above, MR is an analytic technique that has been used to estimate causal relationships between risk factors and exposures and here serves as a good illustration of the possible complications of analytical power and structure. Population stratification can essentially be thought of as the reintroduction of confounding of the genetic instrument (used to proxy the exposure) and disease outcome, thus violating the MR assumption that there is no confounding between the genetic instrument and the outcome. Therefore, GWASs that do not fully account for any ancestral population structure can lead to population stratification (59), and estimates from MR analyses based on the results of that GWAS can potentially be biased by the coincidence of associations among genotype, population structure, and health.

The key question, therefore, is how pervasive this type of structure is and whether it can be demonstrated [as in the study by Haworth et al. (39)]. A recent GWAS of coronavirus disease 2019 (COVID-19) outcomes—substantively aided by UKBB data—is a real-time exemplar of just such potential complications. COVID-19 is an infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which has had a profound impact on the health and lives of people worldwide (103). UKBB has been uniquely positioned to contribute to research on the COVID-19 pandemic. For example, the entire UKBB cohort (approximately 500,000 participants) was invited to receive a self-test kit to determine whether they have SARS-CoV-2 antibodies due to past infection (rather than vaccination). UKBB is also one of the largest contributors to the

COVID-19 Host Genetics Initiative (21; **http://www.covid19hg.org**), an international consortium that has brought together investigators from around the world to investigate the genetic determinants of COVID-19 susceptibility, severity, and outcome. A GWAS conducted by the initiative identified at least 15 genome-wide significant loci associated with increased COVID-19 susceptibility and severity, including variants in or near several immune genes and the *ABO* locus, which determines ABO blood groups. Many of these loci overlap with previously reported associations with lung-related phenotypes or autoimmune/inflammatory diseases, although some loci have no obvious candidate gene (22). Such discoveries not only contribute to global knowledge of the biology of SARS-CoV-2 infection but also provide genetic evidence for drug targets and drug repurposing and help in the development of genetically informed risk assessments of COVID-19 susceptibility. The publicly available GWAS results for COVID-19 susceptibility and severity have also enabled MR studies to evaluate the causal effects of various exposures on COVID-19 outcomes. Indeed, as of December 2021, 59 MR studies had been conducted (**Table 1**).

In an effort to shed light on potential complexities involved in the application of new GWAS results in this way, we evaluated the implications of population structure for UKBB-based GWASs of COVID-19 susceptibility and severity. We did this by considering the association between birth location and PRSs for COVID-19 susceptibility and severity (see the **Supplemental Methods**). First, a PRS representing the aggregate estimated common genetic contributions to COVID-19 susceptibility was associated with birth location in a model that was then adjusted for genotyping array and study center (**Figure 1**). These associations were attenuated in models that incorporated adjustment for 40 genetic principal components that were able to capture structure in the available genetic data; however, this was not always the case for all COVID-19 outcomes. By contrast, for COVID-19 severity (case-only analysis), associations between PRS and location were actually more pronounced in models incorporating adjustment for 40 genetic principal components, potentially reflecting the potentially biasing effects of association analysis within stratified samples (in this case, COVID-19-only participants).

Supplemental Material >

Outside of the new challenge presented by SARS-CoV-2 and COVID-19, a second exemplar lies in the notion that one cannot assume that the availability of big omics data and big GWAS data brings an increased ability to deconstruct networks of complex biological association; indeed, while there will be an ability to discover genetic association signals with good analytical power, redundancy and complexity can impede direct interpretation. A good example comes from the analysis of high-throughput metabolomic data. Metabolomic profiles are the result of genetic and nongenetic factors, provide a readout of biological processes, and can functionally link genetic loci to disease risk factors and disease outcomes (9, 61, 122). Metabolomics technologies based on mass spectrometry and nuclear magnetic resonance (NMR) have enabled the systematic quantification of hundreds of metabolites (the metabolome) from a single biological sample. The analysis of metabolites has enabled more thorough exploration of an individual's metabolic status, offering new opportunities to improve our understanding of the molecular mechanisms underlying human traits and diseases (50). Over the last decade, several metabolite GWASs have been performed (25, 35, 47, 51, 71, 74, 93, 94, 104, 110) to characterize the genetic architecture of blood metabolite variation, and these studies have provided an estimate of the heritability of multiple metabolites and insights into the biological and clinical relevance of these genetic associations (107, 110). Recently, metabolic biomarker data quantified using NMR in approximately 121,000 participants have been made available from UKBB (48, 97). The availability of this large-scale omics measurement combined with genome-wide data has maximized the power to discover genetic loci for a given metabolite and provided a better understanding of the genetic architecture of blood metabolites.

**Table 1 COVID-19 Mendelian randomization studies**

| PubMed ID | Reference | Authors | Year | Exposure | Conclusion |
|---|---|---|---|---|---|
| 34402426 | 2 | Anisul et al. | 2021 | Proteome | ABO protein is associated with several COVID-19 phenotypes. |
| 33757497 | 5 | Au Yeung et al. | 2021 | Cardiometabolic factors | Type 2 diabetes and glycemic traits are not associated with COVID-19 outcomes. |
| 33262790 | 6 | Aung et al. | 2020 | Cardiometabolic factors | Higher BMI and LDL cholesterol are associated with susceptibility to COVID-19. |
| 34557504 | 7 | Baranova et al. | 2021 | Expression and methylation quantitative trait loci | Seven protein-coding genes (*TYK2*, *IFNAR2*, *OAS1*, *OAS3*, *XCR1*, *CCR5*, and *MAPT*) are associated with COVID-19 outcomes, two of which are novel risk genes (*CCR5* and *MAPT*, which encodes the tau protein). |
| 34061844 | 13 | Butler-Laporte et al. | 2021 | Biochemical levels | Higher vitamin D levels are not associated with COVID-19 susceptibility, hospitalization, or severity. |
| 33349849 | 14 | Butler-Laporte et al. | 2021 | Biochemical levels | Lower serum ACE levels are not associated with COVID-19 susceptibility. |
| 34502134 | 16 | Cai et al. | 2021 | Host receptors (ACE2, DC-SIGN, and L-SIGN) | Higher expression of DC-SIGN plasma protein is associated with increased COVID-19 risk, but higher expression of ACE2 or L-SIGN is not. |
| 34840730 | 17 | Cecelja et al. | 2021 | Cardiometabolic factors | There is little evidence for a causal association of cardiovascular risk factors or disease with severe COVID-19. |
| 34237774 | 22 | COVID-19 Host Genetics Initiative | 2021 | Cardiometabolic factors, cardiovascular factors, respiratory factors, biochemical levels, lifestyle factors, psychiatric disorders | Smoking and BMI are associated with severe COVID-19, but type 2 diabetes is not. |
| 34246301 | 23 | Cui & Tian | 2021 | Biochemical levels | Serum vitamin D concentration is not associated with COVID-19 susceptibility, severity, or hospitalization. |
| 34032881 | 26 | Du et al. | 2021 | Cardiometabolic factors, cardiovascular factors, respiratory factors, biochemical levels, lifestyle factors | Comorbidity of inflammatory bowel disease and elevated levels of CRP and IL10 causally increase risk of acute respiratory distress syndrome, while vitamin D supplementation and vasodilators reduce risk. |
| 33714028 | 29 | Fadista et al. | 2021 | Respiratory factors | Idiopathic pulmonary fibrosis is associated with increased risk of severe COVID-19. |
| 34068824 | 30 | Fan et al. | 2021 | Lifestyle factors | There is little evidence that alcohol consumption is associated with risk of SARS-CoV-2 infection in participants either with or without obesity. |
| 33631142 | 32 | Freuer et al. | 2021 | Cardiometabolic factors | BMI is associated with COVID-19 susceptibility and hospitalization. |

(*Continued*)

**Table 1 (*Continued*)**

| PubMed ID | Reference | Authors | Year | Exposure | Conclusion |
|---|---|---|---|---|---|
| 33837377 | 36 | Gaziano et al. | 2021 | Drugs | ACE2, IFNAR2, and IL10RB are associated with COVID-19 hospitalization. |
| 33391794 | 37 | Gill et al. | 2020 | Drugs, biochemical levels | Genetically proxied serum ACE2 levels are not associated with COVID-19 hospitalization. |
| 33604698 | 41 | Hernández Cordero et al. | 2021 | Cardiovascular factors, respiratory factors | Plasma ABO protein is associated with increased risk of severe COVID-19. |
| 33667465 | 43 | Hilser et al. | 2021 | Cardiometabolic factors | Higher HDL cholesterol levels are not associated with decreased risk of COVID-19 susceptibility or mortality. |
| 34866576 | 46 | Huang et al. | 2021 | Cardiometabolic factors, drugs | Higher *HMGCR* expression is associated with increased risk of COVID-19 hospitalization. |
| 33445938 | 56 | Kopeček & Höschl | 2020 | Biochemical levels | Genetically proxied serum vitamin D levels are not associated with COVID-19 infection. |
| 33214204 | 57 | Larsson et al. | 2021 | Immune system | Genetically proxied IL6R inhibition is associated with decreased risk of COVID-19 susceptibility and hospitalization. |
| 33661905 | 60 | Leong et al. | 2021 | Cardiometabolic factors | BMI is associated with COVID-19 severity and hospitalization but not with increased risk of testing positive. |
| 34768390 | 62 | Li et al. | 2021 | Educational attainment | Higher educational attainment is associated with decreased risk of severe COVID-19. It also decreases the risk of COVID-19 hospitalization, but the association is attenuated after adjustment for beta estimates of intelligence in multivariable analysis. Higher intelligence is associated with decreased risk of COVID-19 hospitalization, but the association is attenuated after adjustment for educational attainment. |
| 34163532 | 63 | Li et al. | 2021 | Immune system | Higher CCL4 levels are associated with decreased risk of COVID-19 hospitalization. |
| 33536004 | 64 | Li & Hua | 2021 | Cardiometabolic factors, lifestyle factors | Smoking and higher BMI are associated with increased risk of severe COVID-19 and hospitalization. Genetically proxied physical activity is associated with decreased risk of severe COVID-19. |
| 34521884 | 65 | Li et al. | 2021 | Biomarkers | Vitamin D levels are causally associated with COVID-19 risk. |
| 34530967 | 67 | Liu et al. | 2021 | Biomarkers | There is little evidence that genetically predicted vitamin D levels are associated with COVID-19 risk. |

(*Continued*)

**Table 1 (*Continued*)**

| PubMed ID | Reference | Authors | Year | Exposure | Conclusion |
|---|---|---|---|---|---|
| 33259846 | 68 | Liu et al. | 2021 | Cardiovascular factors, respiratory factors | ILMN_1765156 and ILMN_1791057 probes for *IFNAR2* are associated with COVID-19 hospitalization. |
| 34227468 | 69 | Liu et al. | 2021 | Lifestyle factors | Smoking is associated with *ACE2* expression, but genetically proxied alcohol consumption is not. |
| 34744839 | 70 | Liu et al. | 2021 | Psychiatric disorders | ADHD is associated with increased risk of COVID-19 hospitalization. Genetically predicted COVID-19 severity is significantly associated with schizophrenia. |
| 33846372 | 73 | Lorincz-Comi & Zhu | 2021 | Cardiometabolic factors | Type 2 diabetes and genetically proxied pulse pressure are not associated with COVID-19 hospitalization. |
| 33833219 | 75 | Luykx & Lin | 2021 | Psychiatric disorders | Schizophrenia and Alzheimer's disease are associated with COVID-19 susceptibility. Genetic liability to schizophrenia and bipolar disorder are associated with risk of severe COVID-19. |
| 34553760 | 81 | Ong et al. | 2022 | Gastroesophageal reflux disease | Susceptibility to gastroesophageal reflux disease increases risk of severe COVID-19, COVID-19 hospitalization, and overall risk of COVID-19. |
| 33307546 | 82 | Pairo-Castineira et al. | 2021 | Biochemical levels | Low expression of *IFNAR2* and high expression of *TYK2* are associated with life-threatening COVID-19 disease. High expression of *CCR2* is associated with severe COVID-19. |
| 34478452 | 84 | Papadopoulou et al. | 2021 | COVID-19 susceptibility and severity | COVID-19 risk is associated with increased risk of phlebitis and thrombophlebitis. COVID-19 susceptibility is associated with blood clots in the leg and increased risk of blood clots in the lungs. |
| 32966752 | 85 | Ponsford et al. | 2020 | Cardiometabolic factors, lifestyle factors | Lifetime smoking and higher BMI are associated with increased risk of severe COVID-19, COVID-19 hospitalization, and COVID-19 respiratory failure. |
| 34530029 | 87 | Qiu et al. | 2022 | Metabolite biomarkers | Three blood metabolite concentrations (citrate, M.VLDL.C, and S.VLDL.P) are associated with increased risk of COVID-19 (increasing the COVID-19 infection rate by 24%, 13%, and 10%, respectively), but none of them passed the adjusted significance threshold. |

(*Continued*)

**Table 1 (*Continued*)**

| PubMed ID | Reference | Authors | Year | Exposure | Conclusion |
|---|---|---|---|---|---|
| 34467987 | 88 | Ran et al. | 2021 | COVID-19 susceptibility, cardiometabolic factors, immune system | Genetically proxied COVID-19 is associated with decreased eosinophil cell count, LDL cholesterol, total cholesterol, BCAP29, and KIR2DL5A. |
| 34863930 | 89 | Ran et al. | 2021 | Bone mineral density | Total body bone mineral density is associated with severe COVID-19 in individuals over 60 years old. |
| 34308962 | 90 | Rao et al. | 2021 | Lifestyle factors | Increasing consumption of cigarettes per day increases susceptibility to severe COVID-19 and COVID-19 hospitalization, but alcohol consumption does not. |
| 32430459 | 91 | Rao et al. | 2020 | Cardiometabolic factors, cardiovascular factors, drugs | Type 2 diabetes is associated with increased *ACE2* expression. |
| 33548839 | 95 | Richardson et al. | 2021 | Cardiometabolic factors | Glycoprotein 130 is associated with severe COVID-19. |
| 34127963 | 98 | Roh et al. | 2021 | Cardiovascular factors | *ADAMTS13* is associated with myocardial injury in COVID-19 patients. |
| 34711921 | 99 | Rosoff et al. | 2021 | Lifestyle factors | There is strong evidence that smoking increases the risk of COVID-19 and other respiratory infections even after accounting for other substance use behaviors and cardiometabolic diseases. |
| 33809027 | 111 | Sun et al. | 2021 | Immune system | Higher basophil count, basophil percentage of white blood cells, and myeloid white blood cells are associated with decreased risk of severe COVID-19. Basophil count, basophil percentage of white blood cells, and myeloid white blood cells are not associated with COVID-19 susceptibility. |
| 34428710 | 113 | Tan et al. | 2021 | COVID-19 susceptibility | Genetically predicted COVID-19 is significantly positively associated with hypertension disorders in pregnancy. |
| 34122505 | 119 | Wang et al. | 2021 | Cardiovascular factors, biochemical levels, immune system | Higher neutrophil, monocyte, and lymphocyte counts are associated with decreased risk of COVID-19. |
| 34202464 | 120 | Wang et al. | 2021 | Cardiovascular factors | Blood pressure is associated with severe COVID-19 with respiratory failure. |
| 34304048 | 121 | Wang et al. | 2021 | Telomere length | Shorter telomere length is associated with increased risk of adverse COVID-19 outcomes independent of several major risk factors for COVID-19, including age. |
| 34150709 | 124 | Yoshikawa & Asaba | 2021 | Lifestyle factors | Higher educational attainment is associated with reduced risk of severe COVID-19. |

(*Continued*)

**Table 1 (*Continued*)**

| PubMed ID | Reference | Authors | Year | Exposure | Conclusion |
|---|---|---|---|---|---|
| 34774031 | 125 | Yoshikawa et al. | 2021 | Cardiometabolic factors | Serum ApoB and LDL cholesterol levels are not significantly associated with COVID-19 risk. Higher serum triglyceride levels are suggestively associated with increased risk of COVID-19 susceptibility and hospitalization and are significantly associated with COVID-19 severity. |
| 34496635 | 127 | Zhang et al. | 2021 | Cardiometabolic factors | Higher total cholesterol and ApoB levels might increase the risk of COVID-19 infection. |
| 34189540 | 128 | Zhang et al. | 2021 | Immune system | Failure of NKG2D-mediated activation is associated with severe COVID-19. |
| 33312507 | 129 | Zhang et al. | 2020 | Lifestyle factors | Moderate to vigorous physical activity is not associated with COVID-19 outcomes. |
| 34774005 | 130 | Zhao & Schooling | 2021 | Kidney disorders | Higher estimated glomerular filtration rate is associated with reduced risk of severe COVID-19 but not with risk of COVID-19 hospitalization or infection. Genetically instrumented urine albumin-to-creatinine ratio is not associated with COVID-19 susceptibility or outcomes. |
| 33633408 | 132 | Zhou et al. | 2021 | Biochemical levels | Increased OAS1 levels are associated with reduced risk of COVID-19 susceptibility, hospitalization, ventilation, and death. |
| 34099622 | 133 | Zhou et al. | 2021 | Biochemical levels, immune system | VWF, VWF activity, and ADAMTS13 are associated with COVID-19 severity. |
| 34608450 | 134 | Zhu et al. | 2021 | Hematopoietic traits | White blood cell counts and cholesterol levels are causally associated with COVID-19 severity. |
| 34755518 | 136 | Zuber et al. | 2021 | COVID-19 susceptibility and severity | Critical COVID-19 is associated with an increased risk of ischemic stroke. |

Abbreviations: ACE, angiotensin-converting enzyme; ADHD, attention deficit hyperactivity disorder; ADAMTS13, a disintegrin and metalloproteinase with a thrombospondin type 1 motif, member 13; ApoB, apolipoprotein B; BCAP29, B cell receptor–associated protein 29; BMI, body mass index; CCL4, C-C motif chemokine ligand 4; COVID-19, coronavirus disease 2019; CRP, C-reactive protein; DC-SIGN, dendritic cell–specific intercellular adhesion molecule 3–grabbing nonintegrin; HDL, high-density lipoprotein; IFNAR2, interferon alpha and beta receptor subunit 2; IL6R, interleukin 6 receptor; IL10, interleukin 10; IL10RB, interleukin 10 receptor subunit beta; KIR2DL5A, killer cell immunoglobulin-like receptor 2DL5A; L-SIGN, liver/lymph node–specific intercellular adhesion molecule 3–grabbing nonintegrin; LDL, low-density lipoprotein; M.VLDL.C, total cholesterol in medium very-low-density lipoprotein; NKG2D, natural killer group 2D; OAS1, 2′-5′-oligoadenylate synthetase 1; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2; S.VLDL.P, concentration of small very-low-density lipoprotein particles; VWF, von Willebrand factor.

However, the genetic architecture of blood metabolites is complicated by the high correlation structure and shared biology of the metabolites, which cause complexities when analyzing the causal associations between individual metabolites and disease outcomes using MR analyses. This was exemplified in a recent study that demonstrated that genetic instruments associated with metabolites are likely to be highly pleiotropic, with few SNPs found to be associated with specific metabolites (38). Furthermore, there was a high degree of pleiotropy for metabolite-associated

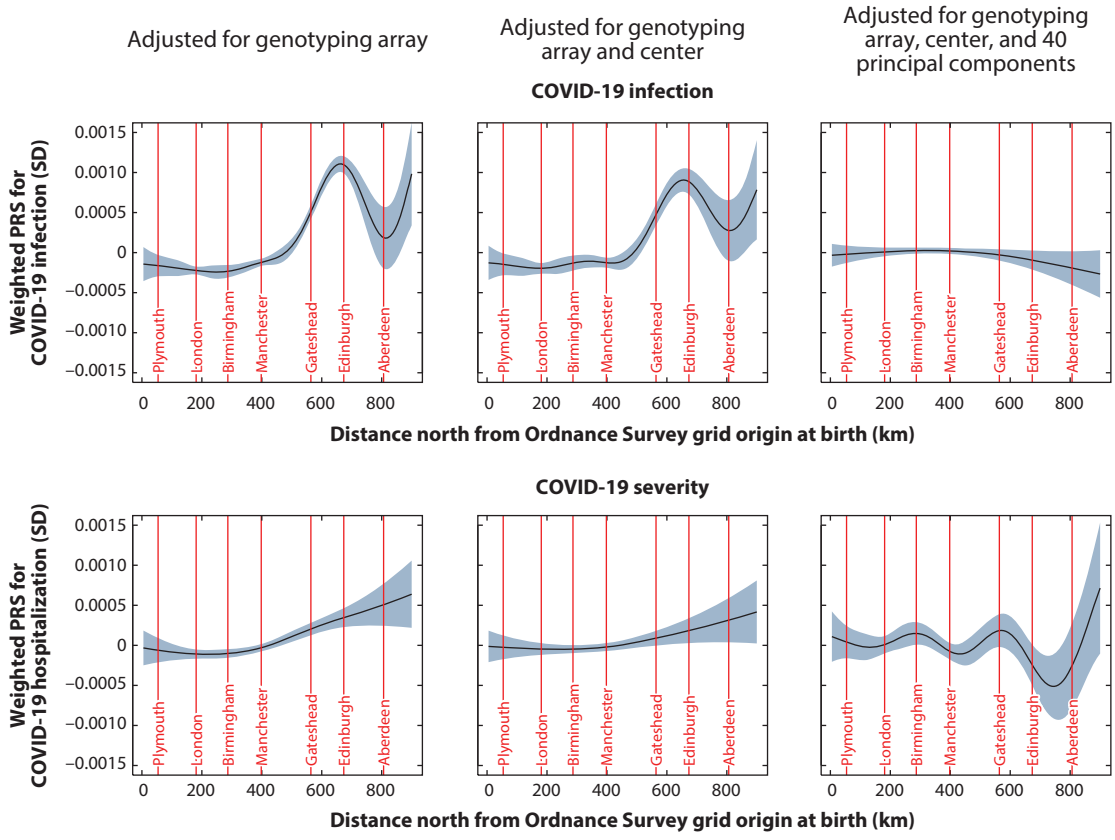


**Figure 1**

Fitted spline regression plots showing the nonlinear distribution of weighted PRSs for COVID-19 susceptibility and severity in a model after adjustment for genotyping array (*left*), adjustment for genotyping array and study center (*middle*), and adjustment for genotyping array, study center, and 40 principal components (*right*). The centers of major cities are marked for reference; shaded areas represent 95% confidence intervals. Abbreviations: COVID-19, coronavirus disease 2019; PRS, polygenic risk score; SD, standard deviation.

SNPs with modifiable risk factors and other disease endpoints. As most metabolites have only a small number of instruments, the use of statistical methods aiming to correct for these biases [e.g., MR-Egger (10) and MR-PRESSO (Mendelian randomization pleiotropy residual sum and outlier) (117)] is not possible, nor is the use of techniques designed to evaluate the effect of multiple correlated exposures [e.g., multivariable MR (100, 101)].

To explore this type of complexity further, we undertook a simple analysis that sought to demonstrate the number of metabolomic features associated with genetic variants in UKBB at a predefined and stringent threshold (see the **Supplemental Methods**). By using recently available NMR data within UKBB to perform a basic GWAS for circulating metabolites, we found that only a few of the plentiful collection of potential genetic associations that would satisfy the conditions to be used as instruments within MR analyses [i.e., (*a*) the genetic variant is associated with the exposure, (*b*) there is no association between the genetic variant and the outcome, and (*c*) the genetic variant is independent of any measured or unmeasured confounding factors] were associated with a specific metabolite. By contrast, numerous loci showed high levels of multimetabolite association, with a median of 34 metabolites associated with each locus (**Figure 2**). The profound

Supplemental Material >

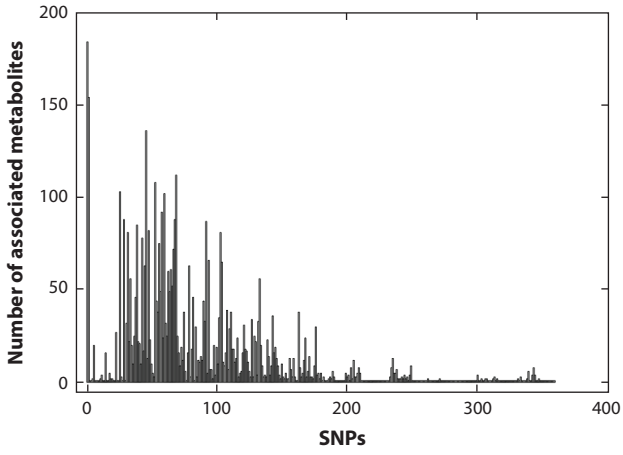


**Figure 2**

Distribution of associations for NMR metabolite instruments (defined as $p < 5 \times 10^{-8}$ and clumped using $R^2 < 0.001$) from GWASs using UKBB data Abbreviations: GWAS, genome-wide association study; NMR, nuclear magnetic resonance; SNP, single-nucleotide polymorphism; UKBB, UK Biobank.

overlap of association signals across metabolites is clearly a complicating factor, and one that would potentially violate assumptions made in analyses such as MR. That is not to say that these associations are uninformative or that these issues are insurmountable; rather, the associations are clear markers of the potential issues that need to be considered when power and precision are able to generate strong association profiles. For studies investigating whether metabolites could be biological pathways relevant to disease onset, a potential way to solve this problem is to conduct profile comparison analysis to examine the overlap between the metabolomic profile of prospective disease risk and that of the risk factor of the disease (e.g., BMI) to identify biological pathways relevant to disease onset (38).

## DISCUSSION

UKBB is a shining example of the impact of large, open access population biobanks in increasing the power to understand the genetic architecture of common traits and diseases. Among a wider set of potential benefits (not all of which have been considered here), the dramatic expansion of GWAS sample sizes improves the power to estimate effect sizes, genomic prediction, and the potential for applied analyses, such as those relating to causal inference. At the same time, however, the availability of substantial analytical power and enabling analytical capacity can increase the complications and inferential complexity associated with any one specific analysis. For example, as described here and in previous studies (20, 39), population structure within the UKBB data could potentially bias association results or their interpretation. The presence of population structure is challenging, requiring methods that are specific to the analytical context and trait. If not properly corrected for, the sampling structure can generate properties in data that can lead to biased inference. Caution is therefore needed in the interpretation of GWAS results using data from UKBB, particularly for loci that demonstrate strong residual associations with birth location even after adjustment for population stratification (20).

Despite this and other limitations mentioned here and elsewhere, UKBB remains an extraordinary resource. In its large data collection, research output, enabling capacity, and likely future

contributions, it has undeniably shaped the modern GWAS era. Most of the problems noted in the analysis of results from UKBB are, and likely will continue to be, a consequence of the misinterpretation of results generated from the UKBB sampling frame, not the sampling frame itself. Used for appropriate analyses, and with the results interpreted in the context of the specific nature of the UKBB data set, UKBB will undoubtedly continue to be a light in the field of human GWASs.

## DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

## ACKNOWLEDGMENTS

N.J.T. is a Wellcome Trust Investigator (202802/Z/16/Z), is the principal investigator of the Avon Longitudinal Study of Parents and Children (MRC and Wellcome Trust 217065/Z/19/Z), is supported by the University of Bristol NIHR Biomedical Research Centre (BRC-1215-2001) and the MRC IEU (MC_UU_00011/1), and works within the Cancer Research UK Integrative Cancer Epidemiology Programme (C18281/A29019). V.Y.T. is supported by the Cancer Research UK Integrative Cancer Epidemiology Programme (C18281/A29019).

## LITERATURE CITED

1. Allison DB, Thiel B, St. Jean P, Elston RC, Infante MC, Schork NJ. 1998. Multiple phenotype modeling in gene-mapping studies of quantitative traits: power advantages. *Am. J. Hum. Genet.* 63:1190–201
2. Anisul M, Shilts J, Schwartzentruber J, Hayhurst J, Buniello A, et al. 2021. A proteome-wide genetic investigation identifies several SARS-CoV-2-exploited host targets of clinical relevance. *eLife* 10:e69719
3. Aragam KG, Chaffin M, Levinson RT, McDermott G, Choi SH, et al. 2019. Phenotypic refinement of heart failure in a national biobank facilitates genetic discovery. *Circulation* 11:489–501
4. Astle WJ, Elding H, Jiang T, Allen D, Ruklisa D, et al. 2016. The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell* 167:1415–29.e19
5. Au Yeung SL, Zhao JV, Schooling CM. 2021. Evaluation of glycemic traits in susceptibility to COVID-19 risk: a Mendelian randomization study. *BMC Med.* 19:72
6. Aung N, Khanji MY, Munroe PB, Petersen SE. 2020. Causal inference for genetic obesity, cardiometabolic profile and COVID-19 susceptibility: a Mendelian randomization study. *Front. Genet.* 11:586308
7. Baranova A, Cao H, Zhang F. 2021. Unraveling risk genes of COVID-19 by multi-omics integrative analyses. *Front. Med.* 8:738687
8. Baselmans BML, Jansen R, Ip HF, van Dongen J, Abdellaoui A, et al. 2019. Multivariate genome-wide analyses of the well-being spectrum. *Nat. Genet.* 51:445–51
9. Blasco H, Nadal-Desbarats L, Pradat PF, Gordon PH, Madji Hounoum B, et al. 2016. Biomarkers in amyotrophic lateral sclerosis: combining metabolomic and clinical parameters to define disease progression. *Eur. J. Neurol.* 23:346–53
10. Bowden J, Davey Smith G, Burgess S. 2015. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* 44:512–25
11. Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, et al. 2015. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* 47:1236–41
12. Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, et al. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447:661–78
13. Butler-Laporte G, Nakanishi T, Mooser V, Morrison DR, Abdullah T, et al. 2021. Vitamin D and COVID-19 susceptibility and severity in the COVID-19 Host Genetics Initiative: a Mendelian randomization study. *PLOS Med.* 18:e1003605

14. Butler-Laporte G, Nakanishi T, Mooser V, Renieri A, Amitrano S, et al. 2021. The effect of angiotensin-converting enzyme levels on COVID-19 susceptibility and severity: a Mendelian randomization study. *Int. J. Epidemiol.* 50:75–86
15. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, et al. 2018. The UK Biobank resource with deep phenotyping and genomic data. *Nature* 562:203–9
16. Cai G, Du M, Bossé Y, Albrecht H, Qin F, et al. 2021. SARS-CoV-2 impairs dendritic cells and regulates DC-SIGN gene expression in tissues. *Int. J. Mol. Sci.* 22:9228
17. Cecelja M, Lewis CM, Shah AM, Chowienczyk P. 2021. Cardiovascular health and risk of hospitalization with COVID-19: a Mendelian randomization study. *JRSM Cardiovasc. Dis.* 10:20480040211059374
18. Chavali S, Barrenas F, Kanduri K, Benson M. 2010. Network properties of human disease genes with pleiotropic effects. *BMC Syst. Biol.* 4:78
19. Conroy M, Sellors J, Effingham M, Littlejohns TJ, Boultwood C, et al. 2019. The advantages of UK Biobank's open-access strategy for health research. *J. Intern. Med.* 286:389–97
20. Cook JP, Mahajan A, Morris AP. 2020. Fine-scale population structure in the UK Biobank: implications for genome-wide association studies. *Hum. Mol. Genet.* 29:2803–11
21. COVID-19 Host Genet. Init. 2020. The COVID-19 Host Genetics Initiative, a global initiative to elucidate the role of host genetic factors in susceptibility and severity of the SARS-CoV-2 virus pandemic. *Eur. J. Hum. Genet.* 28:715–18
22. COVID-19 Host Genet. Init. 2021. Mapping the human genetic architecture of COVID-19. *Nature* 600:472–77
23. Cui Z, Tian Y. 2021. Using genetic variants to evaluate the causal effect of serum vitamin D concentration on COVID-19 susceptibility, severity and hospitalization traits: a Mendelian randomization study. *J. Transl. Med.* 19:300
24. Davey Smith G, Ebrahim S. 2003. 'Mendelian randomization': Can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol.* 32:1–22
25. Draisma HHM, Pool R, Kobl M, Jansen R, Petersen AK, et al. 2015. Genome-wide association study identifies novel genetic variants contributing to variation in blood metabolite levels. *Nat. Commun.* 6:7208
26. Du M, Garcia JGN, Christie JD, Xin J, Cai G, et al. 2021. Integrative omics provide biological and clinical insights into acute respiratory distress syndrome. *Intensive Care Med.* 47:761–71
27. Elliott LT, Sharp K, Alfaro-Almagro F, Shi S, Miller KL, et al. 2018. Genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nature* 562:210–16
28. Elsworth B, Lyon M, Alexander T, Liu Y, Matthews P, et al. 2020. The MRC IEU OpenGWAS data infrastructure. bioRxiv 2020.08.10.244293. **https://doi.org/10.1101/2020.08.10.244293**
29. Fadista J, Kraven LM, Karjalainen J, Andrews SJ, Geller F, et al. 2021. Shared genetic etiology between idiopathic pulmonary fibrosis and COVID-19 severity. *EBioMedicine* 65:103277
30. Fan X, Liu Z, Poulsen KL, Wu X, Miyata T, et al. 2021. Alcohol consumption is associated with poor prognosis in obese patients with COVID-19: a Mendelian randomization study using UK Biobank. *Nutrients* 13:1592
31. Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, et al. 2007. A common variant in the *FTO* gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* 316:889–94
32. Freuer D, Linseisen J, Meisinger C. 2021. Impact of body composition on COVID-19 susceptibility and severity: a two-sample multivariable Mendelian randomization study. *Metabolism* 118:154732
33. Fry A, Littlejohns TJ, Sudlow C, Doherty N, Adamska L, et al. 2017. Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *Am. J. Epidemiol.* 186:1026–34
34. Galesloot TE, Van Steen K, Kiemeney LALM, Janss LL, Vermeulen SH. 2014. A comparison of multivariate genome-wide association methods. *PLOS ONE* 9:e95923
35. Gallois A, Mefford J, Ko A, Vaysse A, Julienne H, et al. 2019. A comprehensive study of metabolite genetics reveals strong pleiotropy and heterogeneity across time and context. *Nat. Commun.* 10:4788

36. Gaziano L, Giambartolomei C, Pereira AC, Gaulton A, Posner DC, et al. 2021. Actionable druggable genome-wide Mendelian randomization identifies repurposing opportunities for COVID-19. *Nat. Med.* 27:668–76
37. Gill D, Arvanitis M, Carter P, Hernández Cordero AI, Jo B, et al. 2020. ACE inhibition and cardiometabolic risk factors, lung *ACE2* and *TMPRSS2* gene expression, and plasma ACE2 levels: a Mendelian randomization study. *R. Soc. Open Sci.* 7:200958
38. Guida F, Tan VY, Corbin LJ, Smith-Byrne K, Alcala K, et al. 2021. The blood metabolome of incident kidney cancer: a case-control study nested within the MetKid consortium. *PLOS Med.* 18:e1003786
39. Haworth S, Mitchell R, Corbin L, Wade KH, Dudding T, et al. 2019. Apparent latent structure within the UK Biobank sample has implications for epidemiological analysis. *Nat. Commun.* 10:333
40. Hemani G, Zheng J, Elsworth B, Wade KH, Haberland V, et al. 2018. The MR-Base platform supports systematic causal inference across the human phenome. *eLife* 7:e34408
41. Hernández Cordero AI, Li X, Milne S, Yang CX, Bossé Y, et al. 2021. Multi-omics highlights ABO plasma protein as a causal risk factor for COVID-19. *Hum. Genet.* 140:969–79
42. Hill WD, Marioni RE, Maghzian O, Ritchie SJ, Hagenaars SP, et al. 2019. A combined analysis of genetically correlated traits identifies 187 loci and a role for neurogenesis and myelination in intelligence. *Mol. Psychiatry* 24:169–81
43. Hilser JR, Han Y, Biswas S, Gukasyan J, Cai Z, et al. 2021. Association of serum HDL-cholesterol and apolipoprotein A1 levels with risk of severe SARS-CoV-2 infection. *J. Lipid Res.* 62:100061
44. Hou K, Burch KS, Majumdar A, Shi H, Mancuso N, et al. 2019. Accurate estimation of SNP-heritability from biobank-scale data irrespective of genetic architecture. *Nat. Genet.* 51:1244–51
45. Hu Y, Lu Q, Liu W, Zhang Y, Li M, Zhao H. 2017. Joint modeling of genetically correlated diseases and functional annotations increases accuracy of polygenic risk prediction. *PLOS Genet.* 13:e1006836
46. Huang W, Xiao J, Ji J, Chen L. 2021. Association of lipid-lowering drugs with COVID-19 outcomes from a Mendelian randomization study. *eLife* 10:e73873
47. Illig T, Gieger C, Zhai G, Römisch-Margl W, Wang-Sattler R, et al. 2010. A genome-wide perspective of genetic variation in human metabolism. *Nat. Genet.* 42:137–41
48. Julkunen H, Cichońska A, Slagboom PE, Würtz P. 2021. Metabolic biomarker profiling for identification of susceptibility to severe pneumonia and COVID-19 in the general population. *eLife* 10:e63033
49. Kalra G, Milon B, Casella AM, Herb BR, Humphries E, et al. 2020. Biological insights from multi-omic analysis of 31 genomic risk loci for adult hearing difficulty. *PLOS Genet.* 16:e1009025
50. Kastenmüller G, Raffler J, Gieger C, Suhre K. 2015. Genetics of human metabolism: an update. *Hum. Mol. Genet.* 24:R93–101
51. Kettunen J, Demirkan A, Würtz P, Draisma HHM, Haller T, et al. 2016. Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. *Nat. Commun.* 7:11122
52. Keyes KM, Westreich D. 2019. UK Biobank, big data, and the consequences of non-representativeness. *Lancet* 393:1297
53. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, et al. 2018. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* 50:1219–24
54. Khera AV, Chaffin M, Wade KH, Zahid S, Brancale J, et al. 2019. Polygenic prediction of weight and obesity trajectories from birth to adulthood. *Cell* 177:587–96.e9
55. Klei L, Luca D, Devlin B, Roeder K. 2008. Pleiotropy and principal components of heritability combine to increase power for association analysis. *Genet. Epidemiol.* 32:9–19
56. Kopeček M, Höschl C. 2020. Season of the year, vitamin D and COVID-19. *Cas. Lek. Cesk.* 159:312–16
57. Larsson SC, Burgess S, Gill D. 2021. Genetically proxied interleukin-6 receptor inhibition: opposing associations with COVID-19 and pneumonia. *Eur. Respir. J.* 57:2003545
58. Lawlor DA. 2016. Commentary: two-sample Mendelian randomization: opportunities and challenges. *Int. J. Epidemiol.* 45:908–15
59. Lawson DJ, Davies NM, Haworth S, Ashraf B, Howe L, et al. 2020. Is population structure in the genetic biobank era irrelevant, a challenge, or an opportunity? *Hum. Genet.* 139:23–41
60. Leong A, Cole JB, Brenner LN, Meigs JB, Florez JC, Mercader JM. 2021. Cardiometabolic risk factors for COVID-19 susceptibility and severity: a Mendelian randomization analysis. *PLOS Med.* 18:e1003553

61. Lewis GD, Wei R, Liu E, Yang E, Shi X, et al. 2008. Metabolite profiling of blood from individuals undergoing planned myocardial infarction reveals early markers of myocardial injury. *J. Clin. Investig.* 118:3503–12
62. Li GHY, Lam SKK, Wong ICK, Chu JKP, Cheung CL. 2021. Education attainment, intelligence and COVID-19: a Mendelian randomization study. *J. Clin. Med.* 10:4870
63. Li M, Yeung CHC, Schooling CM. 2021. Circulating cytokines and coronavirus disease: a bi-directional Mendelian randomization study. *Front. Genet.* 12:680646
64. Li S, Hua X. 2021. Modifiable lifestyle factors and severe COVID-19 risk: a Mendelian randomisation study. *BMC Med. Genom.* 14:38
65. Li X, van Geffen J, van Weele M, Zhang X, He Y, et al. 2021. An observational and Mendelian randomisation study on vitamin D and COVID-19 risk in UK Biobank. *Sci. Rep.* 11:18262
66. Littlejohns TJ, Sudlow C, Allen NE, Collins R. 2019. UK Biobank: opportunities for cardiovascular research. *Eur. Heart J.* 40:1158–66
67. Liu D, Tian QY, Zhang J, Hou HF, Li Y, et al. 2021. Association between 25 hydroxyvitamin D concentrations and the risk of COVID-19: a Mendelian randomization study. *Biomed. Environ. Sci.* 34:750–54
68. Liu D, Yang J, Feng B, Lu W, Zhao C, Li L. 2021. Mendelian randomization analysis identified genes pleiotropically associated with the risk and prognosis of COVID-19. *J. Infect.* 82:126–32
69. Liu H, Xin J, Cai S, Jiang X. 2021. Mendelian randomization analysis provides causality of smoking on the expression of ACE2, a putative SARS-CoV-2 receptor. *eLife* 10:e64188
70. Liu N, Tan JS, Liu L, Wang Y, Hua L, Qian Q. 2021. Genetic predisposition between COVID-19 and four mental illnesses: a bidirectional, two-sample Mendelian randomization study. *Front. Psychiatry* 12:746276
71. Long T, Hicks M, Yu HC, Biggs WH, Kirkness EF, et al. 2017. Whole-genome sequencing identifies common-to-rare variants associated with human blood metabolites. *Nat. Genet.* 49:568–78
72. Loos RJF, Janssens ACJW. 2017. Predicting polygenic obesity using genetic information. *Cell Metab.* 25:535–43
73. Lorincz-Comi N, Zhu X. 2021. Cardiometabolic risks of SARS-CoV-2 hospitalization using Mendelian randomization. *Sci. Rep.* 11:7848
74. Lotta LA, Pietzner M, Stewart ID, Wittemans LBL, Li C, et al. 2021. A cross-platform approach identifies genetic regulators of human metabolism and health. *Nat. Genet.* 53:54–64
75. Luykx JJ, Lin BD. 2021. Are psychiatric disorders risk factors for COVID-19 susceptibility and severity? A two-sample, bidirectional, univariable, and multivariable Mendelian randomization study. *Transl. Psychiatry* 11:210
76. Maier R, Moser G, Chen GB, Ripke S, Coryell WH, et al. 2015. Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder, and major depressive disorder. *Am. J. Hum. Genet.* 96:283–94
77. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, et al. 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* 9:356–69
78. Neale BM. 2017. Insights from estimates of SNP-heritability for >2,000 traits and disorders in UK Biobank. *Neale Lab*. **http://www.nealelab.is/blog/2017/9/20/insights-from-estimates-of-snp-heritability-for-2000-traits-and-disorders-in-uk-biobank**
79. Neale BM. 2018. UK Biobank. *Neale Lab*. **http://www.nealelab.is/uk-biobank**
80. Neale BM. 2020. Genetic correlation between traits and disorders in the UK Biobank. *Neale Lab*. **https://ukbb-rg.hail.is**
81. Ong J-S, Gharahkhani P, Vaughan TL, Whiteman D, Kendall BJ, MacGregor S. 2022. Assessing the genetic relationship between gastro-esophageal reflux disease and risk of COVID-19 infection. *Hum. Mol. Genet.* 31:471–80
82. Pairo-Castineira E, Clohisey S, Klaric L, Bretherick AD, Rawlik K, et al. 2021. Genetic mechanisms of critical illness in COVID-19. *Nature* 591:92–98
83. Panagiotou OA, Willer CJ, Hirschhorn JN, Ioannidis JPA. 2013. The power of meta-analysis in genome-wide association studies. *Annu. Rev. Genom. Hum. Genet.* 14:441–65

84. Papadopoulou A, Musa H, Sivaganesan M, McCoy D, Deloukas P, Marouli E. 2021. COVID-19 susceptibility variants associate with blood clots, thrombophlebitis and circulatory diseases. *PLOS ONE* 16:e0256988
85. Ponsford MJ, Gkatzionis A, Walker VM, Grant AJ, Wootton RE, et al. 2020. Cardiometabolic traits, sepsis, and severe COVID-19: a Mendelian randomization investigation. *Circulation* 142:1791–93
86. Porter HF, O'Reilly PF. 2017. Multivariate simulation framework reveals performance of multi-trait GWAS methods. *Sci. Rep.* 7:38837
87. Qiu S, Wang D, Zhang Y, Hu Y. 2022. Mendelian randomization reveals potential causal candidates for COVID-19 in 123 blood metabolites. *J. Infect.* 84:248–88
88. Ran S, Su K, Zhang S, Liu B. 2021. The association between coronavirus disease 2019 infection and blood constituents: a Mendelian randomization analysis. *J. Infect. Dis.* 224:922–24
89. Ran S, Zhang S, Chen H, Zhao M, Liu B. 2021. Total body bone mineral density and severe COVID-19: a Mendelian randomization analysis in five age strata. *Bone* 155:116281
90. Rao S, Baranova A, Cao H, Chen J, Zhang X, Zhang F. 2021. Genetic mechanisms of COVID-19 and its association with smoking and alcohol consumption. *Brief. Bioinform.* 22:bbab284. Correction. 2021. *Brief. Bioinform.* 22:bbab357
91. Rao S, Lau A, So HC. 2020. Exploring diseases/traits and blood proteins causally related to expression of ACE2, the putative receptor of SARS-CoV-2: a Mendelian randomization analysis highlights tentative relevance of diabetes-related traits. *Diabetes Care* 43:1416–26
92. Rashkin SR, Graff RE, Kachuri L, Thai KK, Alexeeff SE, et al. 2020. Pan-cancer study detects genetic risk variants and shared genetic basis in two large cohorts. *Nat. Commun.* 11:4423
93. Rhee EP, Ho JE, Chen MH, Shen D, Cheng S, et al. 2013. A genome-wide association study of the human metabolome in a community-based cohort. *Cell Metab.* 18:130–43
94. Rhee EP, Yang Q, Yu B, Liu X, Cheng S, et al. 2016. An exome array study of the plasma metabolome. *Nat. Commun.* 7:12360
95. Richardson TG, Fang S, Mitchell RE, Holmes MV, Davey Smith G. 2021. Evaluating the effects of cardiometabolic exposures on circulating proteins which may contribute to severe SARS-CoV-2. *EBioMedicine* 64:103228
96. Ripatti S, Tikkanen E, Orho-Melander M, Havulinna AS, Silander K, et al. 2010. A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses. *Lancet* 376:1393–400
97. Ritchie SC, Surendran P, Karthikeyan S, Lambert SA, Bolton T, et al. 2021. Quality control and removal of technical variation of NMR metabolic biomarker data in ~120,000 UK Biobank participants. medRxiv 2021.09.24.21264079. **https://doi.org/10.1101/2021.09.24.21264079**
98. Roh J, Kitchen R, Guseh JS, McNeill J, Aid M, et al. 2021. Plasma proteomics of COVID-19 associated cardiovascular complications: implications for pathophysiology and therapeutics. Res. Square 539712. **https://doi.org/10.21203/rs.3.rs-539712/v1**
99. Rosoff DB, Yoo J, Lohoff FW. 2021. Smoking is significantly associated with increased risk of COVID-19 and other respiratory infections. *Commun. Biol.* 4:1230
100. Sanderson E, Davey Smith G, Windmeijer F, Bowden J. 2019. An examination of multivariable Mendelian randomization in the single-sample and two-sample summary data settings. *Int. J. Epidemiol.* 48:713–27. Correction. 2020. *Int. J. Epidemiol.* 49:1057
101. Sanderson E, Spiller W, Bowden J. 2021. Testing and correcting for weak and pleiotropic instruments in two-sample multivariable Mendelian randomization. *Stat. Med.* 40:5434–52
102. Scuteri A, Sanna S, Chen WM, Uda M, Albai G, et al. 2007. Genome-wide association scan shows genetic variants in the *FTO* gene are associated with obesity-related traits. *PLOS Genet.* 3:e115
103. Sharma A, Tiwari S, Deb MK, Marty JL. 2020. Severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2): a global pandemic and treatment strategies. *Int. J. Antimicrob. Agents* 56:106054
104. Shin SY, Fauman EB, Petersen AK, Krumsiek J, Santos R, et al. 2014. An atlas of genetic influences on human blood metabolites. *Nat. Genet.* 46:543–50
105. Sinnott-Armstrong N, Tanigawa Y, Amar D, Mars NJ, Aguirre M, et al. 2019. Genetics of 38 blood and urine biomarkers in the UK Biobank. *Nat. Genet.* 53:185–94
106. Smith JG. 2017. Molecular epidemiology of heart failure: translational challenges and opportunities. *JACC Basic Transl. Sci.* 2:757–69

107. Soininen P, Kangas AJ, Würtz P, Suna T, Ala-Korpela M. 2015. Quantitative serum nuclear magnetic resonance metabolomics in cardiovascular epidemiology and genetics. *Circ. Cardiovasc. Genet.* 8:192–206
108. Spencer CCA, Su Z, Donnelly P, Marchini J. 2009. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLOS Genet.* 5:e1000477
109. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, et al. 2015. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Med.* 12:e1001779
110. Suhre K, Shin SY, Petersen AK, Mohney RP, Meredith D, et al. 2011. Human metabolic individuality in biomedical and pharmaceutical research. *Nature* 477:54–60
111. Sun Y, Zhou J, Ye K. 2021. White blood cells and severe COVID-19: a Mendelian randomization study. *J. Pers. Med.* 11:195
112. Swanson JM. 2012. The UK Biobank and selection bias. *Lancet* 380:110
113. Tan JS, Liu NN, Guo TT, Hu S, Hua L. 2021. Genetic predisposition to COVID-19 may increase the risk of hypertension disorders in pregnancy: a two-sample Mendelian randomization study. *Pregnancy Hypertens.* 26:17–23
114. Timpson NJ, Greenwood CMT, Soranzo N, Lawson DJ, Richards JB. 2018. Genetic architecture: the shape of the genetic contribution to human traits and disease. *Nat. Rev. Genet.* 19:110–24
115. Turley P, Walters RK, Maghzian O, Okbay A, Lee JJ, et al. 2018. Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat. Genet.* 50:229–37
116. van der Sluis S, Verhage M, Posthuma D, Dolan CV. 2010. Phenotypic complexity, measurement bias, and poor phenotypic resolution contribute to the missing heritability problem in genetic association studies. *PLOS ONE* 5:e13929
117. Verbanck M, Chen CY, Neale B, Do R. 2018. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat. Genet.* 50:693–98
118. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, et al. 2017. 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* 101:5–22
119. Wang K, Qu M, Ding L, Shi X, Wang C, et al. 2021. Liver and kidney function biomarkers, blood cell traits and risk of severe COVID-19: a Mendelian randomization study. *Front. Genet.* 12:647303
120. Wang M, Gao J, Liu J, Zhao X, Lei Y. 2021. Genomic association versus serological determination of ABO blood types in a Chinese cohort, with application in Mendelian randomization. *Genes* 12:959
121. Wang Q, Codd V, Raisi-Estabragh Z, Musicha C, Bountziouka V, et al. 2021. Shorter leukocyte telomere length is associated with adverse COVID-19 outcomes: a cohort study in UK Biobank. *EBioMedicine* 70:103485
122. Yazdani A, Yazdani A, Saniei A, Boerwinkle E. 2016. A causal network analysis in an observational study identifies metabolomics pathways influencing plasma triglyceride levels. *Metabolomics* 12:104
123. Yengo L, Sidorenko J, Kemper KE, Zheng Z, Wood AR, et al. 2018. Meta-analysis of genome-wide association studies for height and body mass index in ∼700 000 individuals of European ancestry. *Hum. Mol. Genet.* 27:3641–49
124. Yoshikawa M, Asaba K. 2021. Educational attainment decreases the risk of COVID-19 severity in the European population: a two-sample Mendelian randomization study. *Front. Public Health* 9:673451
125. Yoshikawa M, Asaba K, Nakayama T. 2021. Estimating causal effects of atherogenic lipid-related traits on COVID-19 susceptibility and severity using a two-sample Mendelian randomization approach. *BMC Med. Genom.* 14:269
126. Zeng H, Lin C, Wang S, Zheng Y, Gao X. 2021. Genetically predicted body composition in relation to cardiometabolic traits: a Mendelian randomization study. *Eur. J. Epidemiol.* 36:1157–68
127. Zhang K, Dong SS, Guo Y, Tang SH, Wu H, et al. 2021. Causal associations between blood lipids and COVID-19 risk: a two-sample Mendelian randomization study. *Arterioscler. Thromb. Vasc. Biol.* 41:2802–10
128. Zhang S, Cooper-Knock J, Weimer AK, Harvey C, Julian TH, et al. 2021. Common and rare variant analyses combined with single-cell multiomics reveal cell-type-specific molecular mechanisms of COVID-19 severity. medRxiv 2021.06.15.21258703. **https://doi.org/10.1101/2021.06.15.21258703**
129. Zhang X, Li X, Sun Z, He Y, Xu W, et al. 2020. Physical activity and COVID-19: an observational and Mendelian randomisation study. *J. Glob. Health* 10:020514

130. Zhao JV, Schooling CM. 2021. Using genetics to understand the role of kidney function in COVID-19: a Mendelian randomization study. *BMC Nephrol.* 22:381
131. Zhao Q, Wang J, Spiller W, Bowden J, Small DS. 2019. Two-sample instrumental variable analyses using heterogeneous samples. *Stat. Sci.* 34:317–33
132. Zhou S, Butler-Laporte G, Nakanishi T, Morrison DR, Afilalo J, et al. 2021. A Neanderthal OAS1 isoform protects individuals of European ancestry against COVID-19 susceptibility and severity. *Nat. Med.* 27:659–67
133. Zhou Y, Qian X, Liu Z, Yang H, Liu T, et al. 2021. Coagulation factors and the incidence of COVID-19 severity: Mendelian randomization analyses and supporting evidence. *Signal Transduct. Target. Ther.* 6:222
134. Zhu H, Zheng F, Li L, Jin Y, Luo Y, et al. 2021. A Chinese host genetic study discovered IFNs and causality of laboratory traits on COVID-19 severity. *iScience* 24:103186
135. Zhu W, Zhang H. 2009. Why do we test multiple traits in genetic association studies? *J. Korean Stat. Soc.* 38:1–10
136. Zuber V, Cameron A, Myserlis EP, Bottolo L, Fernandez-Cadenas I, et al. 2021. Leveraging genetic data to elucidate the relationship between COVID-19 and ischemic stroke. *J. Am. Heart Assoc.* 10:e022433